THE COMPUTER SECURITY GROUP AT UC SANTA BARBARA

Δ / Delta

Automatic Identification of Unknown Web-Based Infection Campaigns

Kevin Borgolte kevinbo@cs.ucsb.edu

Christopher Kruegel chris@cs.ucsb.edu

Giovanni Vigna vigna@cs.ucsb.edu

University of California, Santa Barbara

November 4th, 2013

CCS 2013 / Session 1-C / Malware

Motivation

Cybercriminals using Red Kit infect enough sites to increase the number of users who receive malware warnings by **32 million**.

Motivation

Cybercriminals using Red Kit infect enough sites to increase the number of users who receive malware warnings by **32 million**.

A large campaign infects more than **106,000 unique sites in July**, directing people to sites launching the Blackhole Exploit Kit.

Motivation

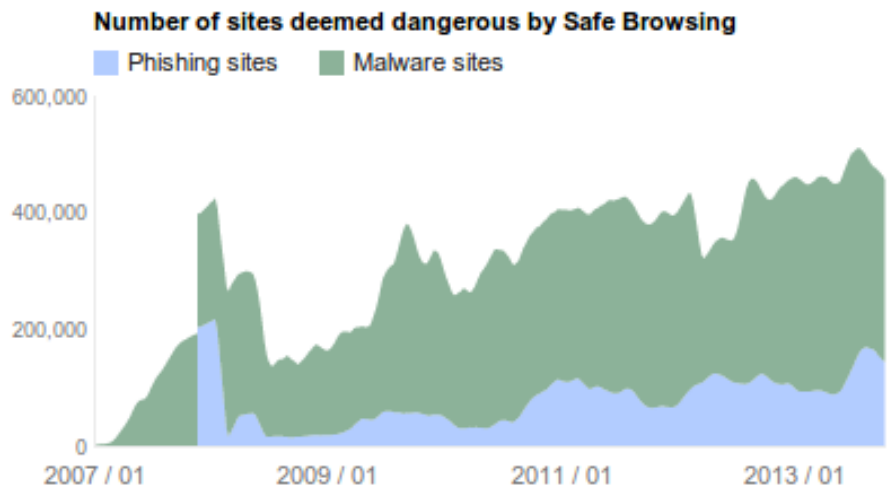
Cybercriminals using Red Kit infect enough sites to increase the number of users who receive malware warnings by **32 million**.

A large campaign infects more than **106,000 unique sites in July**, directing people to sites launching the Blackhole Exploit Kit.

A campaign targeting vulnerabilities in Java and Acrobat Reader infects more than **7,500 sites**. As a result, more than **28.6 million Safe Browsing API users** receive malware warnings **during this week**.

Motivation

- Increasing number of compromised websites each year
- Web being used more and more
- Prior work detects if website is malicious



Infection Vector and Campaign

Infection Vector and Campaign

Infection Vector and Campaign

```
<iframe src='http://wsfgfdgrtyhgfd.net/adv/193/new.php'></iframe>
```

Infection Vector and Campaign

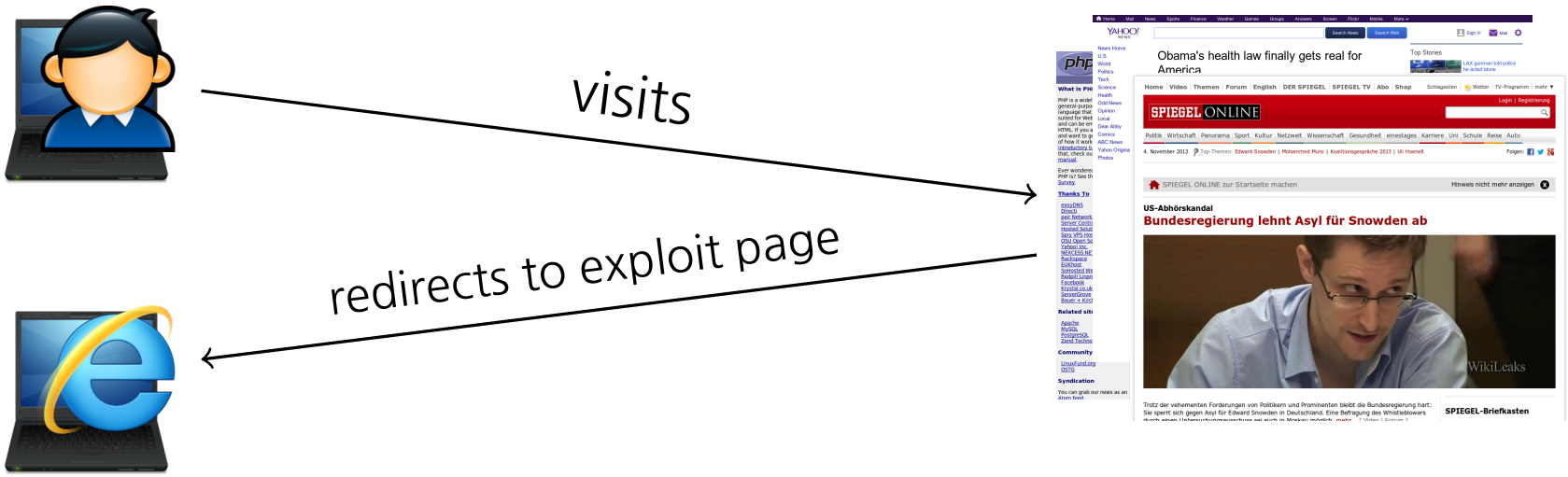
```
<iframe src='http://wsfgfdgrtyhgfd.net/adv/193/new.php'></iframe>
```

Infection Vector and Campaign

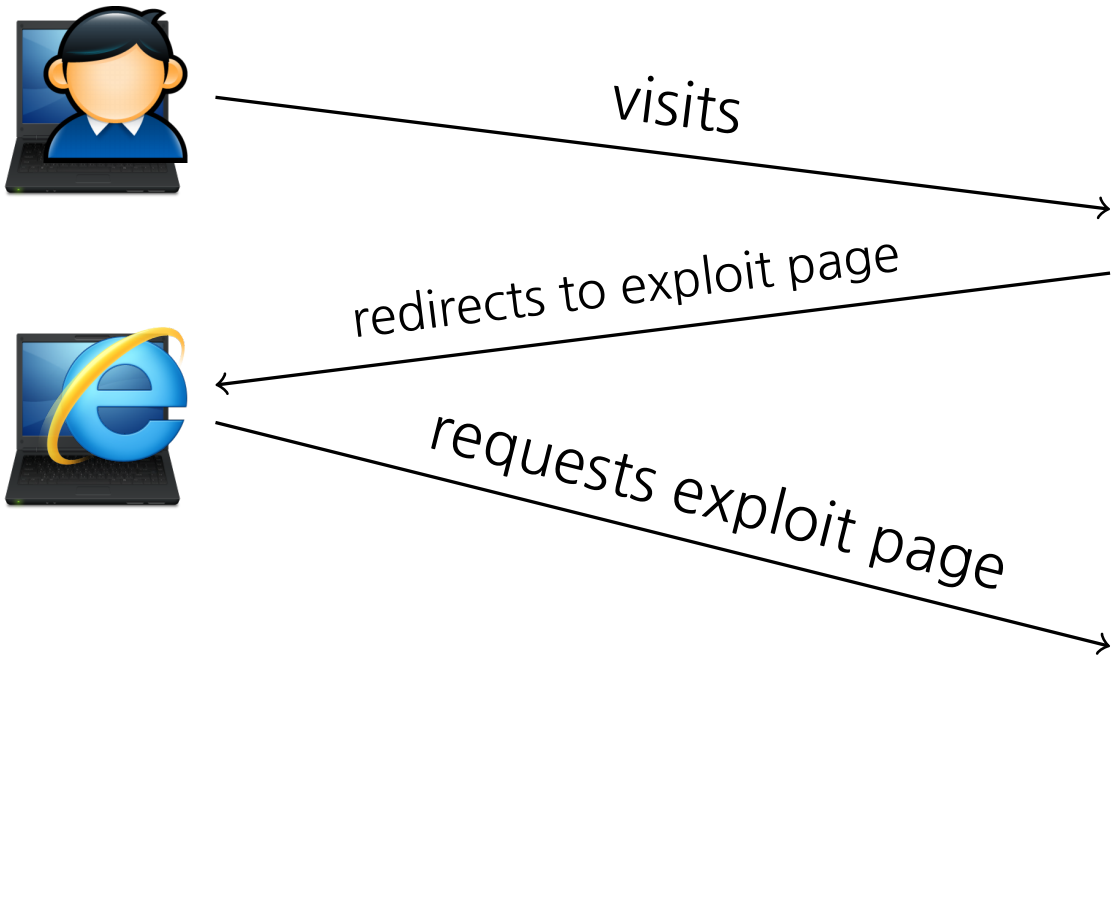
Infection Vector and Campaign

visits

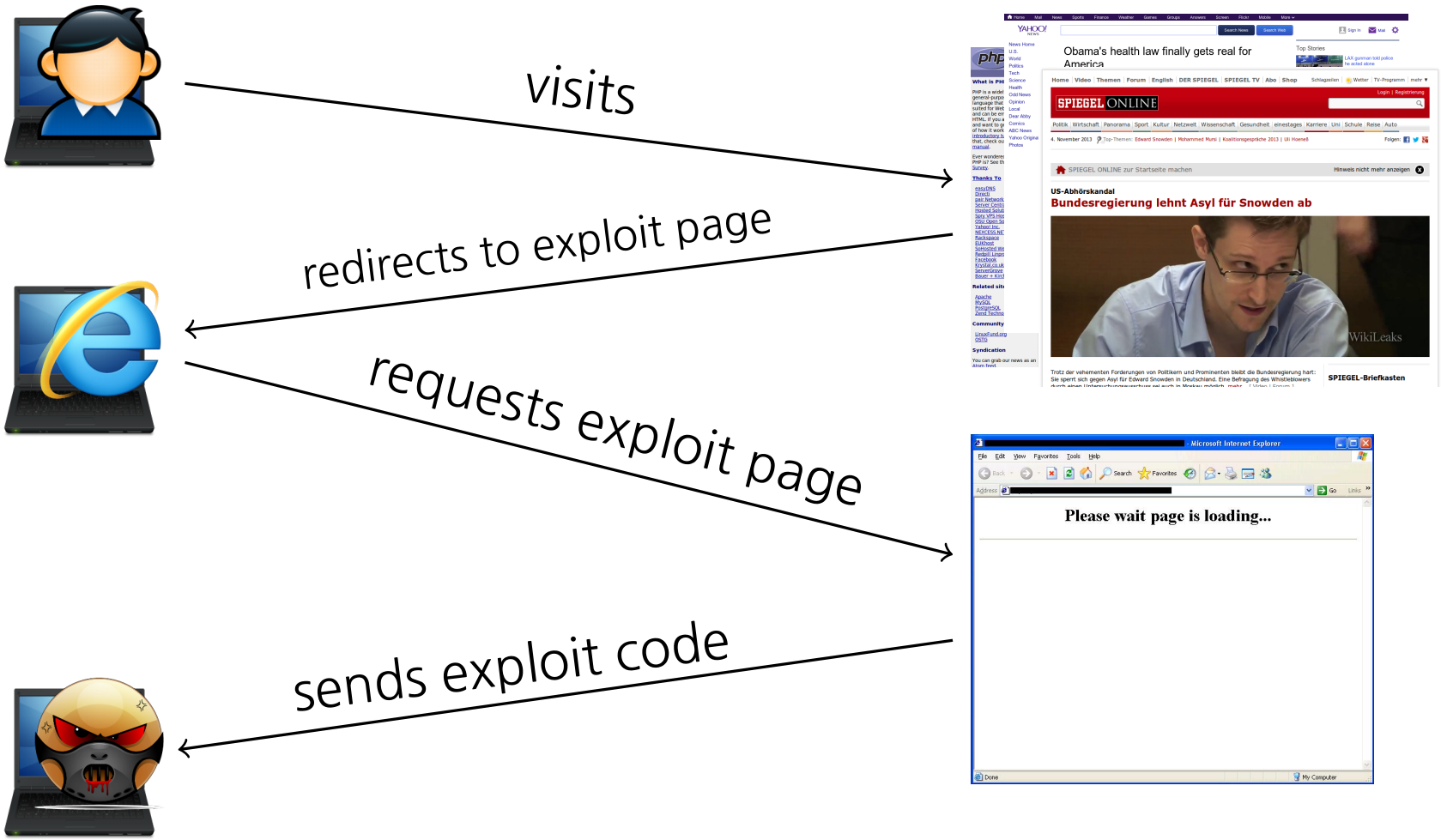
Infection Vector and Campaign



Infection Vector and Campaign



Infection Vector and Campaign



How/why did this website become malicious?

How/why did this website become malicious?

- Websites are being modified

How/why did this website become malicious?

- Websites are being modified
- Identify those modifications
  - Compare to previous version of website

How/why did this website become malicious?

- Websites are being modified
- Identify those modifications
  - Compare to previous version of website
- Cluster similar modifications together

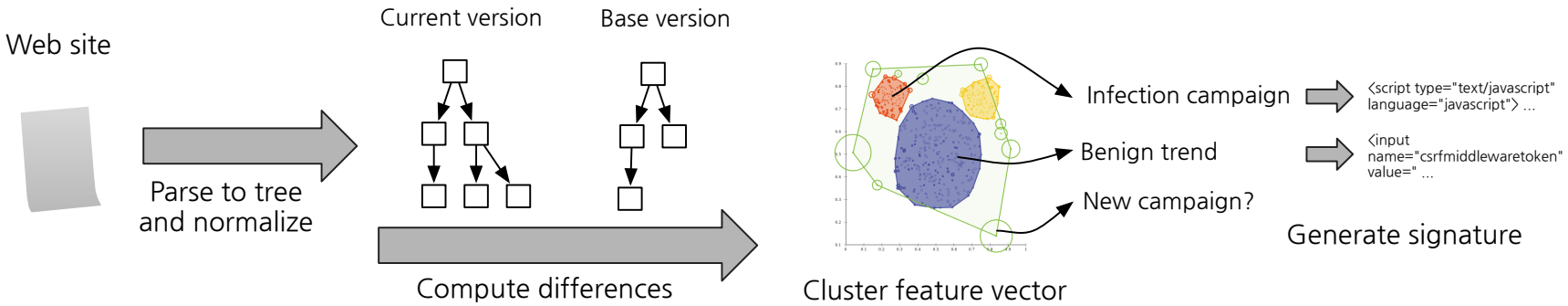
How/why did this website become malicious?

- Websites are being modified
- Identify those modifications
  - Compare to previous version of website
- Cluster similar modifications together
- Analyze if cluster is malicious or not

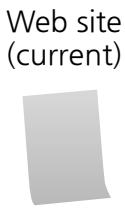
How/why did this website become malicious?

- Websites are being modified
- Identify those modifications
  - Compare to previous version of website
- Cluster similar modifications together
- Analyze if cluster is malicious or not
- Generate signature as model of the campaign

Delta - Approach

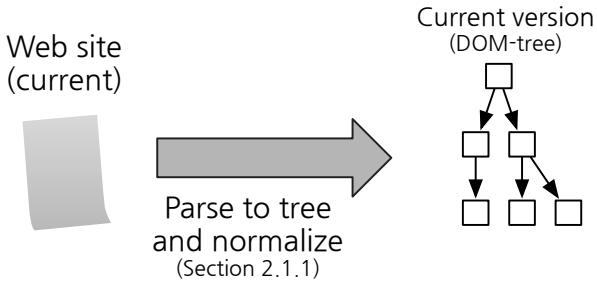


Delta - Approach

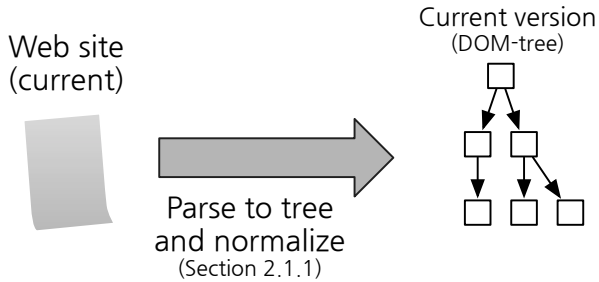


- No client-side script execution
  - Snapshot problem

Delta - Approach

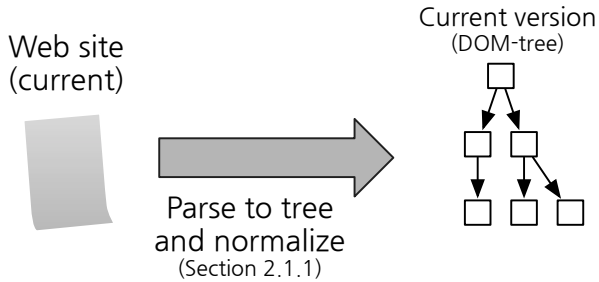


Delta - Approach




```
<a href="http://www.sigsac.org/ccs/CCS2013/" alt='CCS'>CCS</a>
```

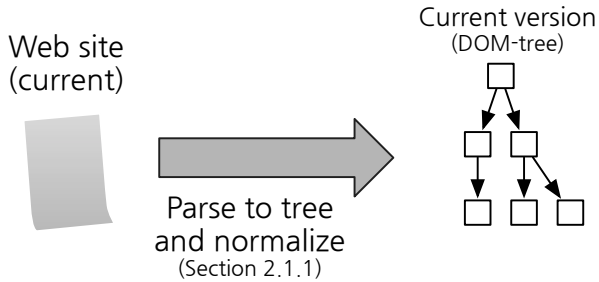
Delta - Approach




```
<a href="http://www.sigsac.org/ccs/CCS2013/" alt='CCS'>CCS</a>
```

```
<a alt='CCS' href='http://www.sigsac.org/ccs/CCS2013/'>CCS</a>
```

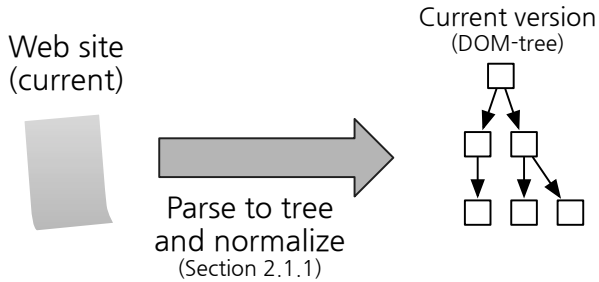
Delta - Approach




```
<a href="http://www.sigsac.org/ccs/CCS2013/" alt='CCS'>CCS</a>

<a alt='CCS' href='http://www.sigsac.org/ccs/CCS2013/'>CCS</a>
```

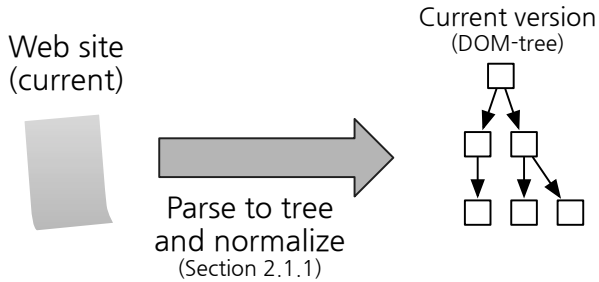
Delta - Approach




```
<a href="http://www.sigsac.org/ccs/CCS2013/" alt='CCS'>CCS</a>

<a alt='CCS' href='http://www.sigsac.org/ccs/CCS2013/'>CCS</a>
```

Delta - Approach



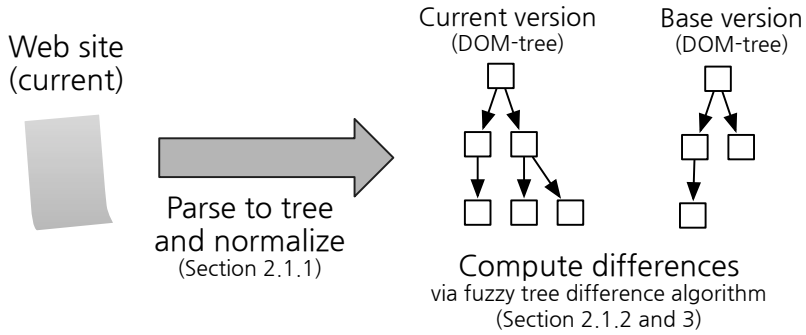

```
<a href="http://www.sigsac.org/ccs/CCS2013/" alt='CCS'>CCS</a>

<a alt='CCS' href='http://www.sigsac.org/ccs/CCS2013/'>CCS</a>
```

Normalization

```
<a alt="CCS" href="http://www.sigsac.org/ccs/CCS2013/">CCS</a>
```

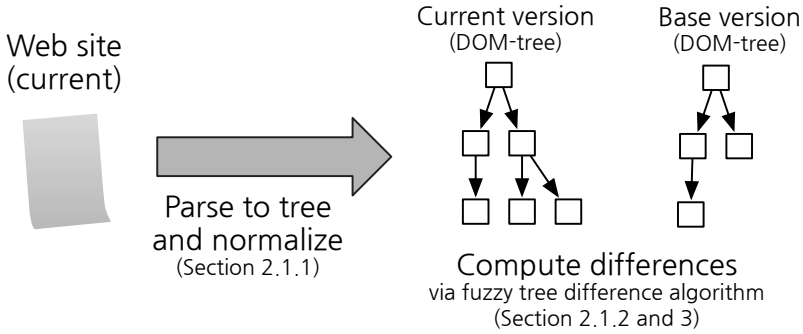
Delta - Approach

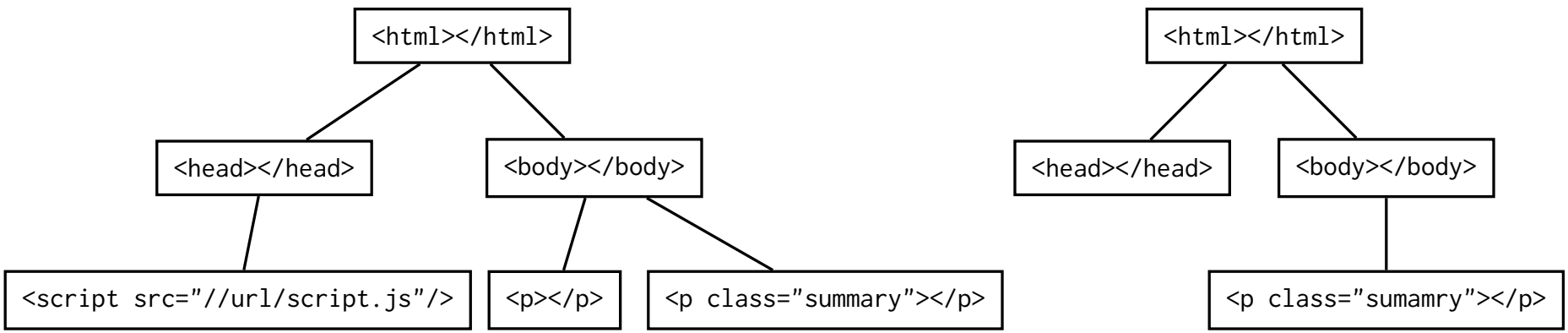


Fuzzy-tree difference:

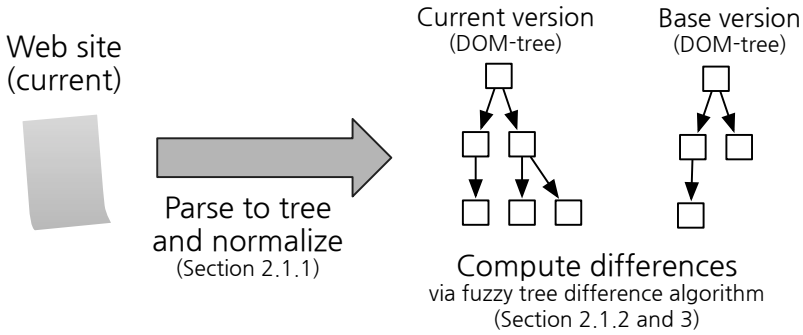
- Tree is considered unordered
- Fuzzy on normalized tags (Jaro distance)

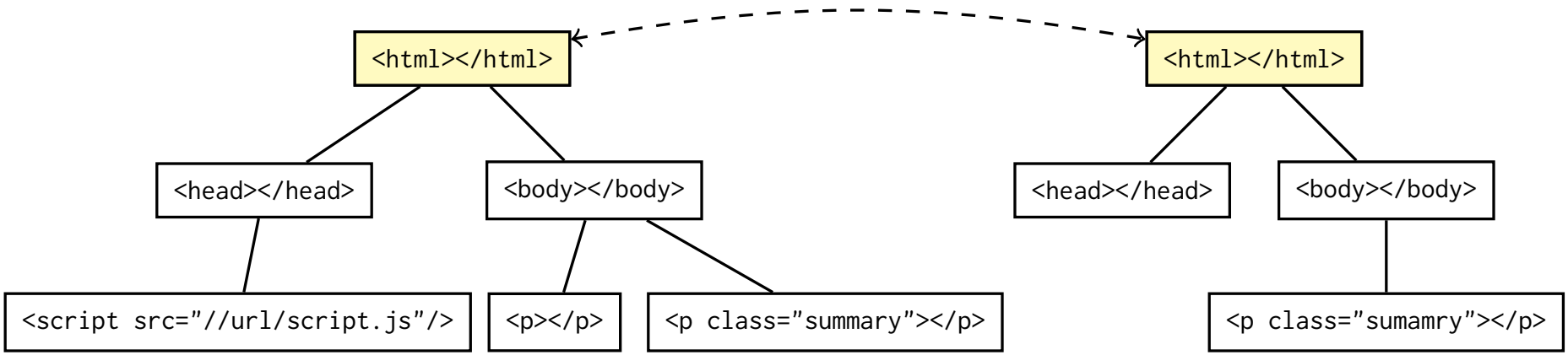
Delta - Approach



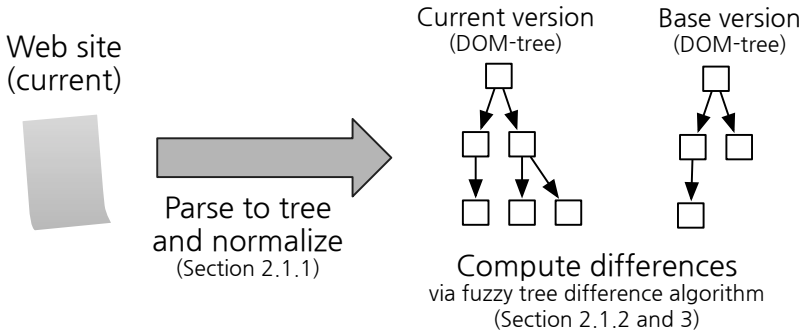


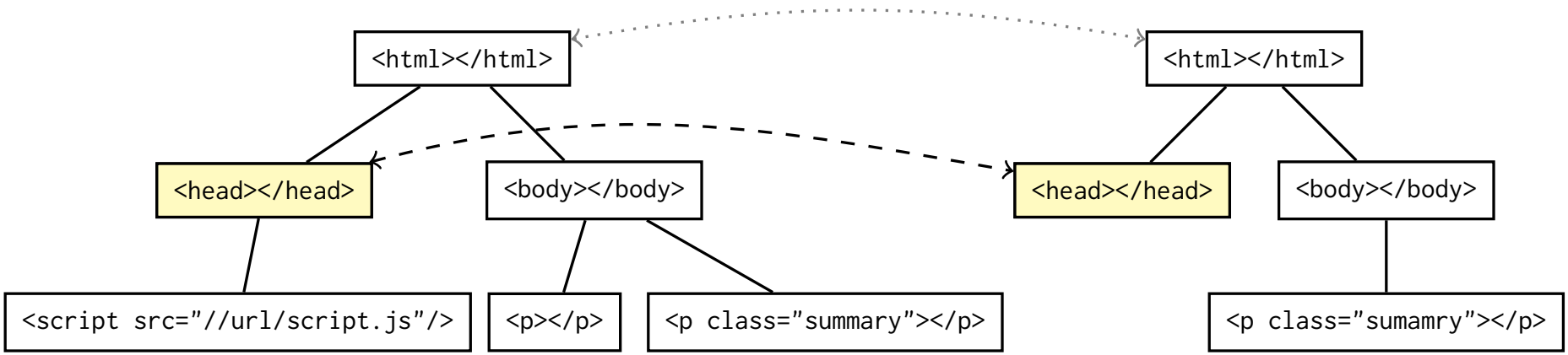
Delta - Approach



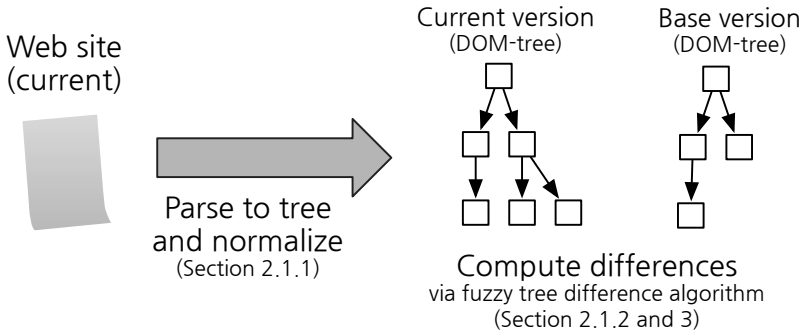


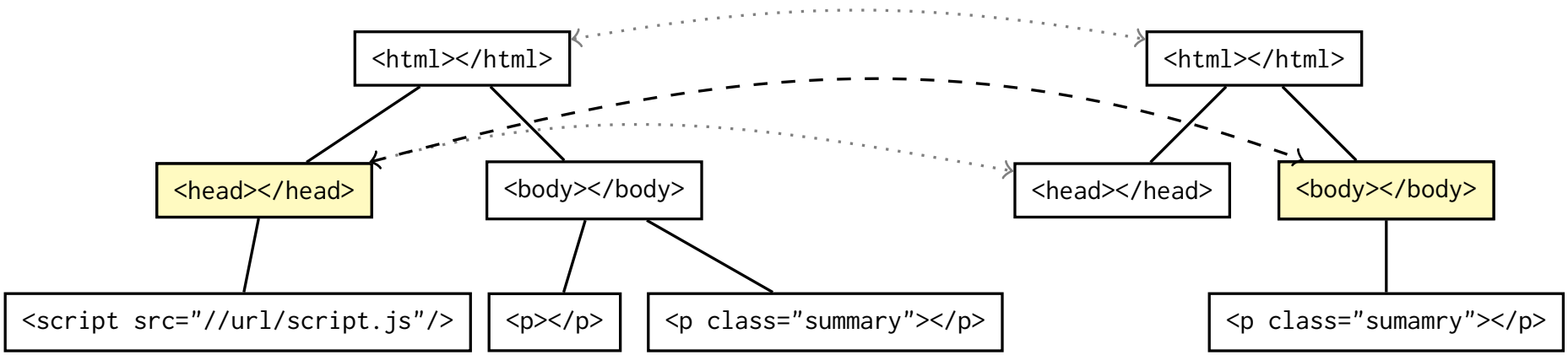
Delta - Approach



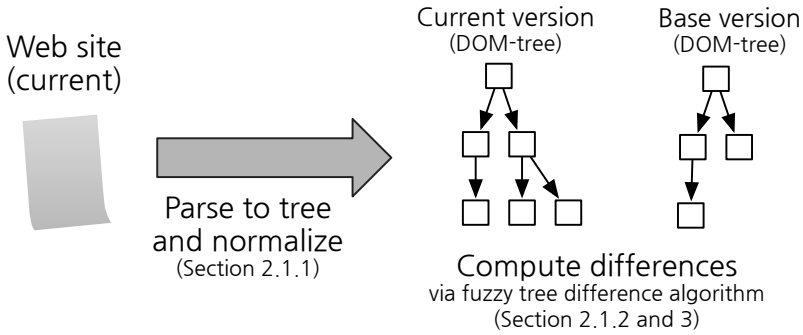


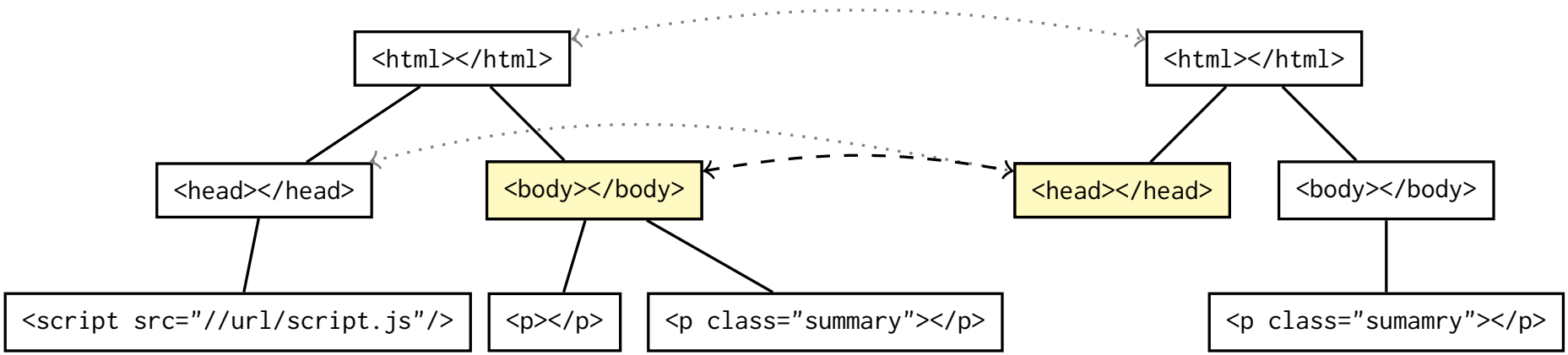
Delta - Approach



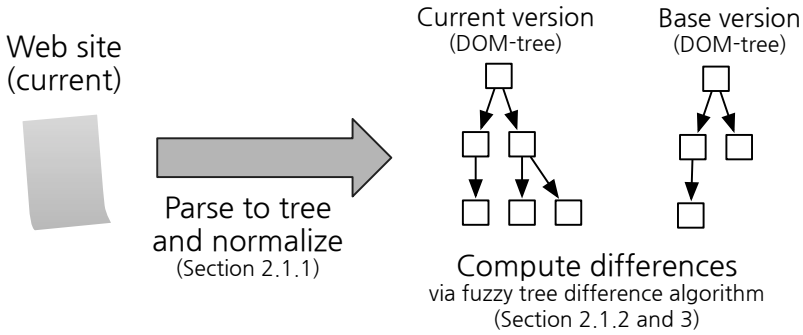


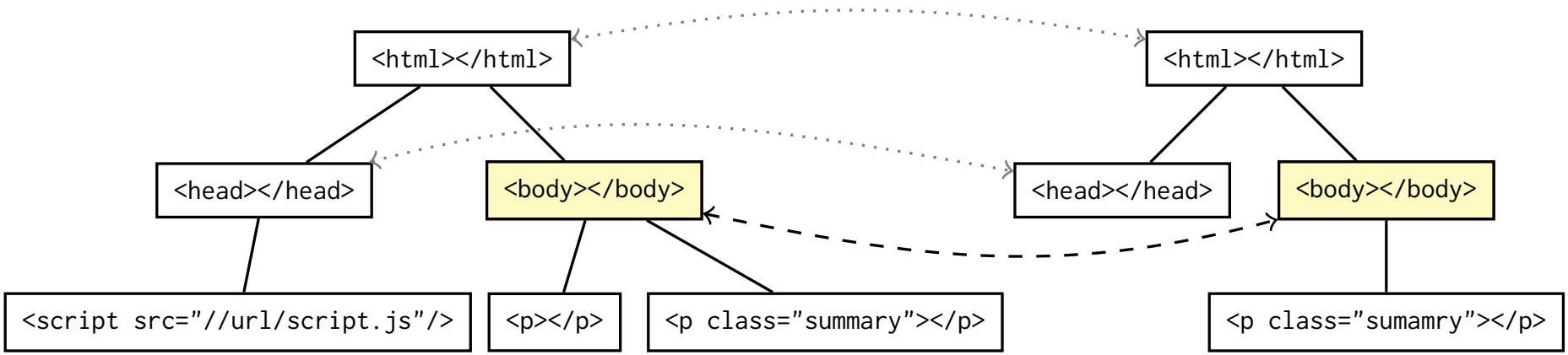
Delta - Approach



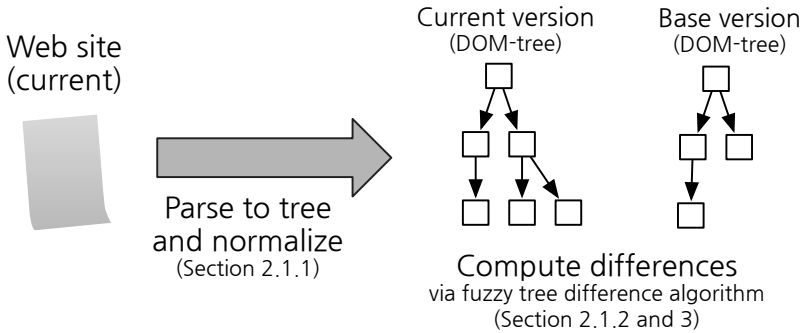


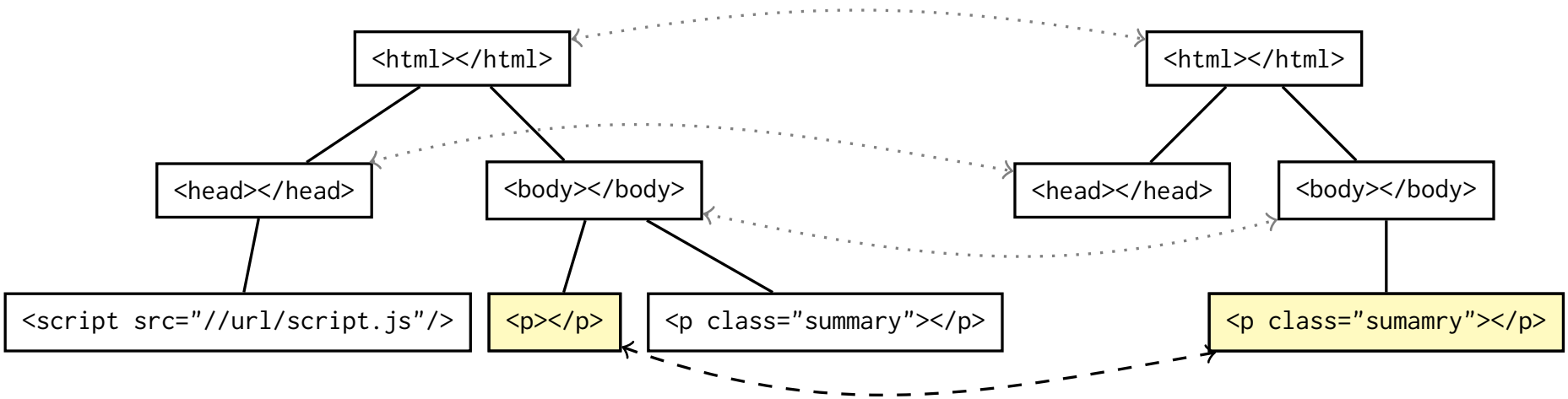
Delta - Approach



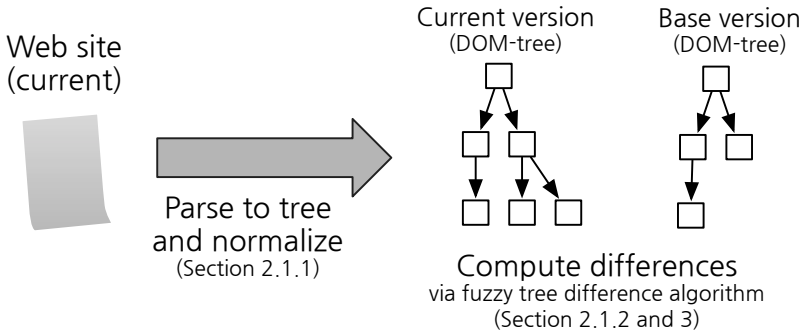


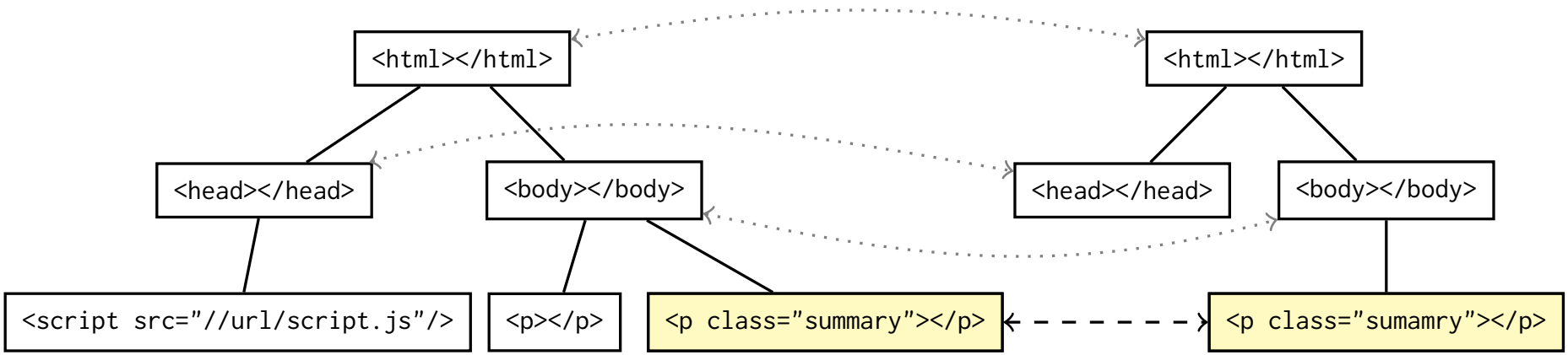
Delta - Approach



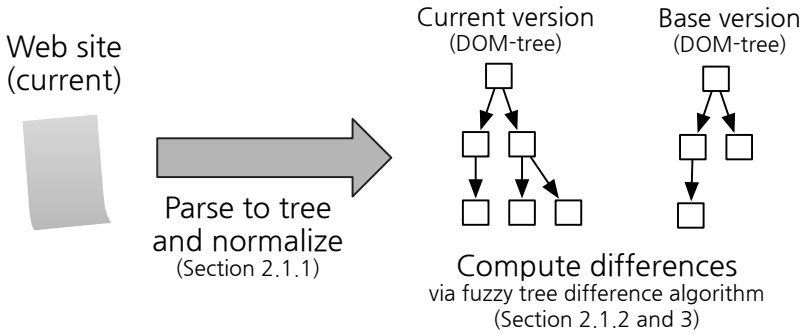


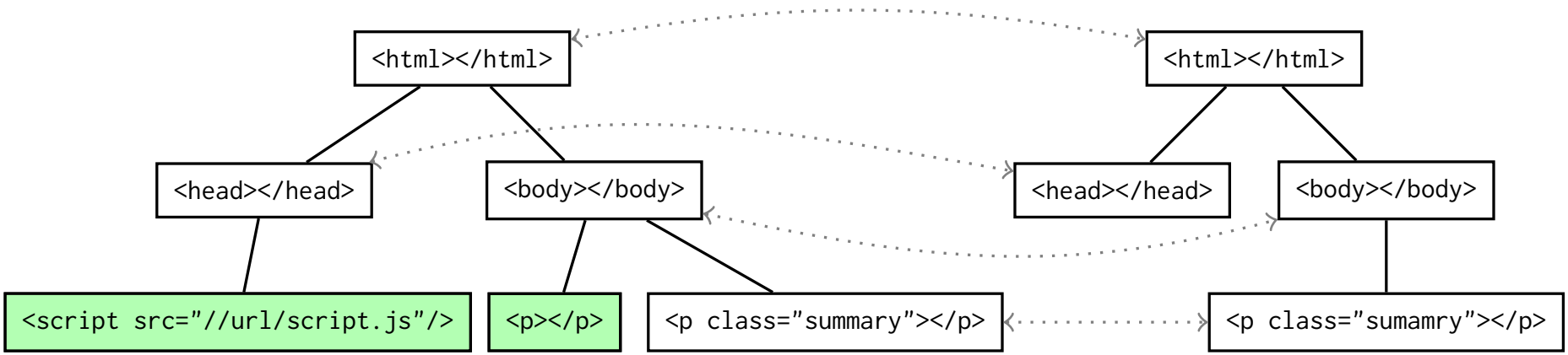
Delta - Approach



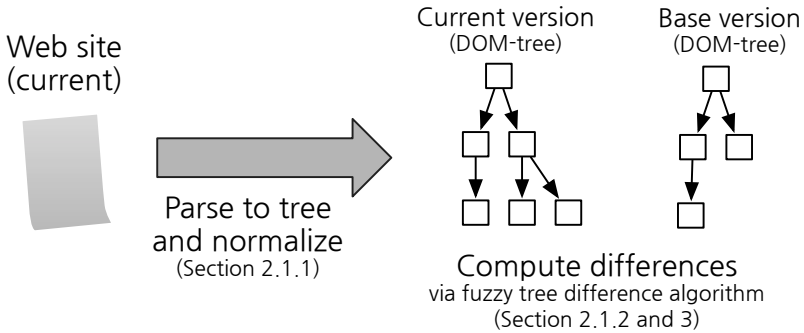


Delta - Approach





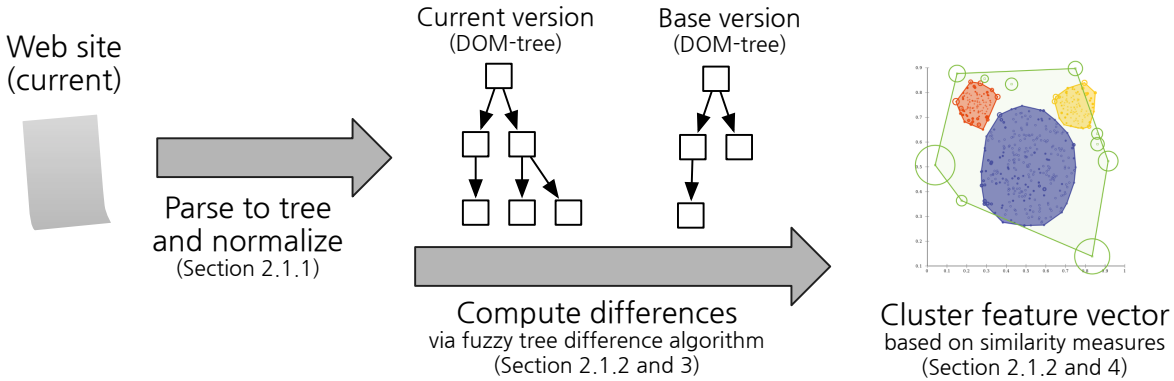
Delta - Approach




```
<script src="//url/script.js"/>
```

```
<p></p>
```

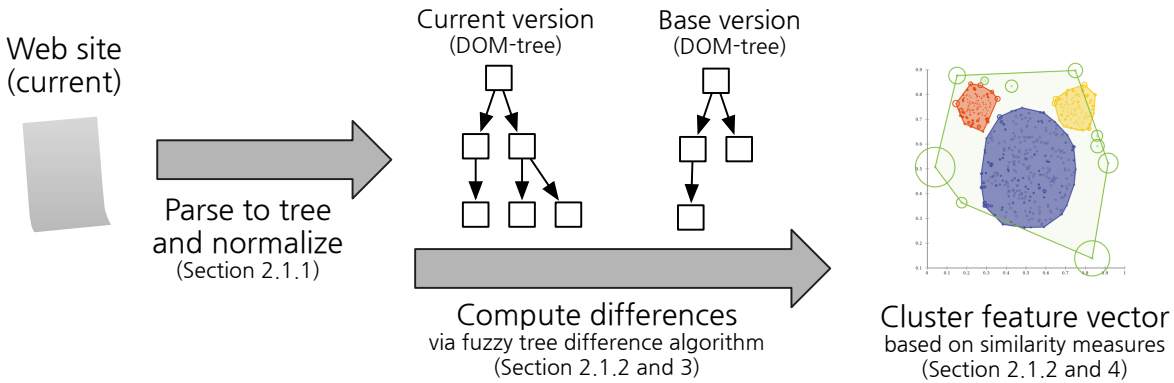
Delta - Approach



Similarity measures

- Template propagation
- Shannon entropy
- Character count/distribution
- Approx. Kolmogorov complexity
- Script inclusion
- …

Delta - Approach

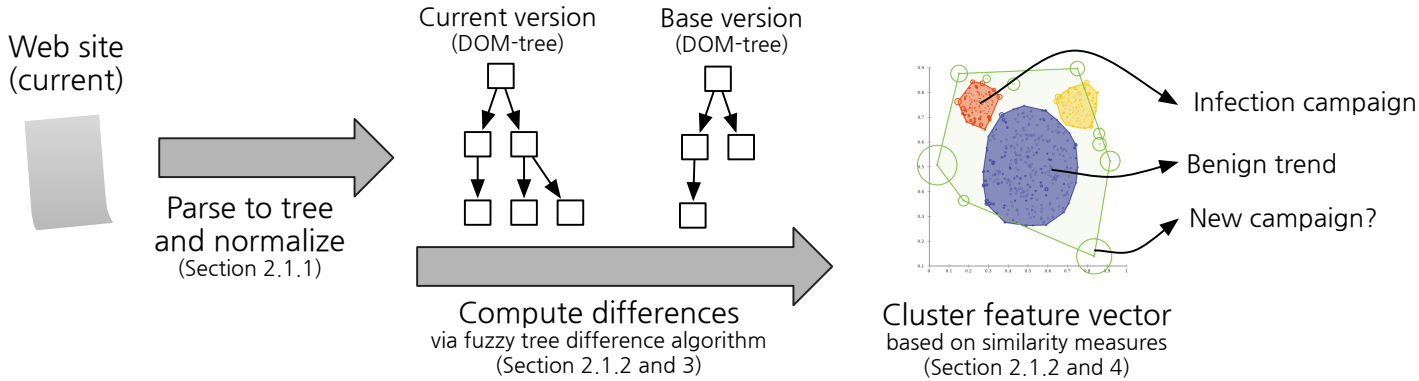


Similarity measures

- Template propagation
- Shannon entropy
- Character count/distribution
- Approx. Kolmogorov complexity
- Script inclusion
- …

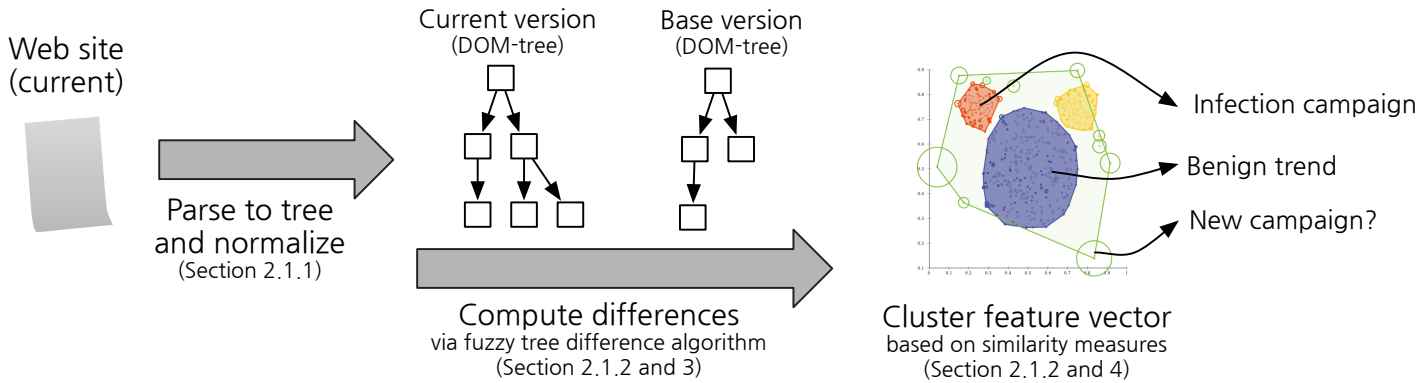
~250 dimensional feature space

Delta - Approach



- Sample from cluster
- Classify behavior for samples
- Assign label to cluster

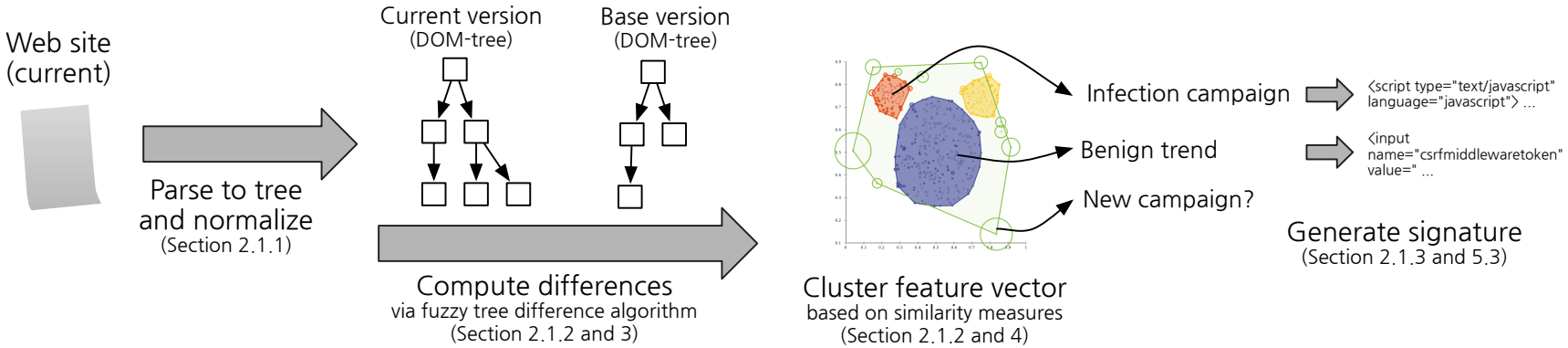
Delta - Approach



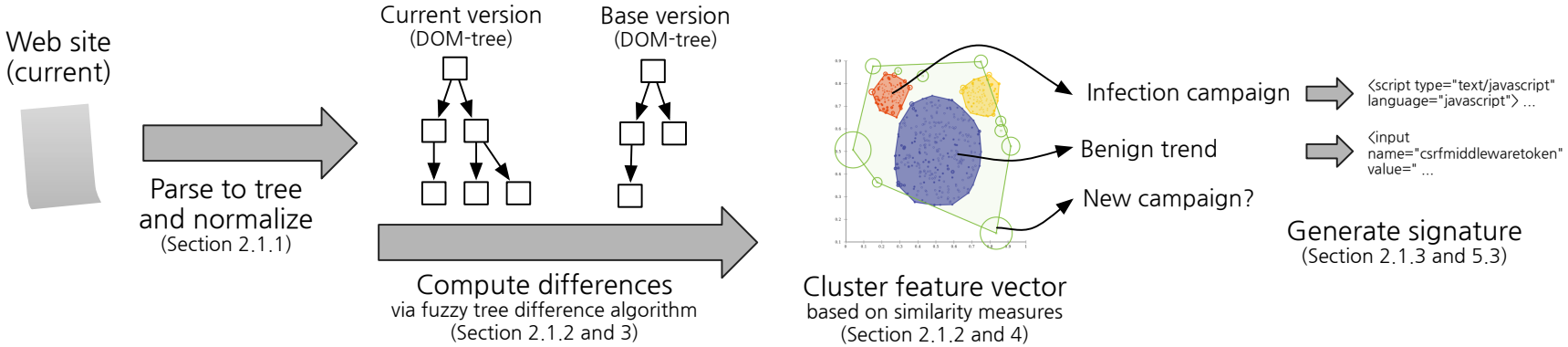
- Sample from cluster
- Classify behavior for samples
- Assign label to cluster

- Density-based clustering
- Outliers acceptable

Delta - Approach



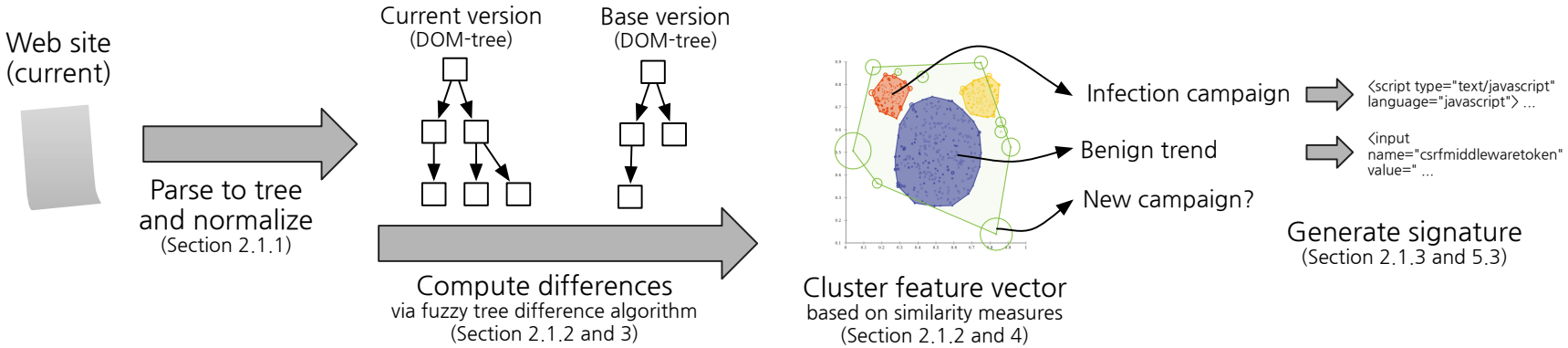
Delta - Approach




```
<script src="http://abc.org/2fcab58712467eab4004583eb8fb7f89.js" />
<script src="http://abc.org/2fcab50712467eab4004583eb8fb7f89.js" />
<script src="http://adc.org/2fcab50712467eab4004583eb8fb7f89.js" />
<script src="http://abc.net/2fcab50712467eab4004583eb8fb7f89.js" />

                                    ...
```

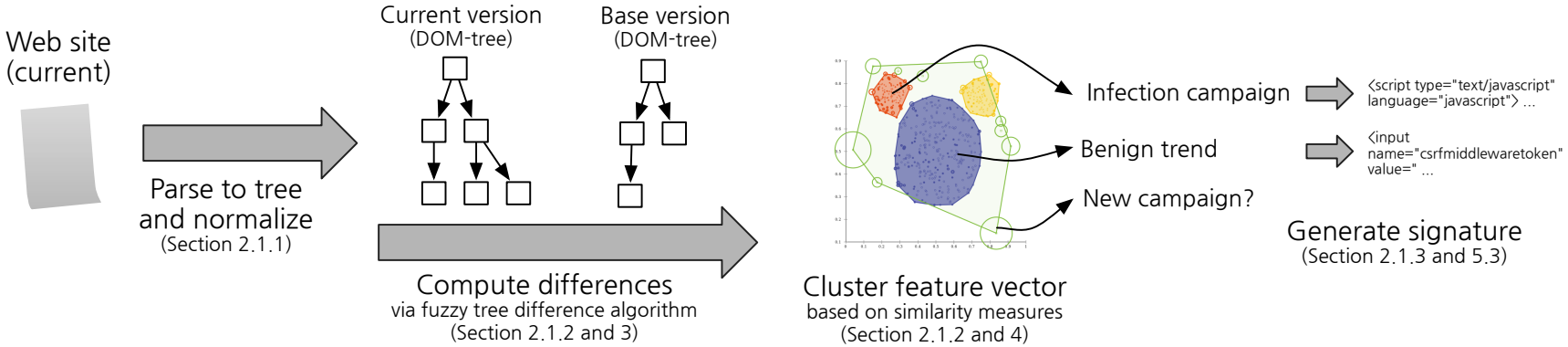
Delta - Approach




```
<script src="http://abc.org/2fcab58712467eab4004583eb8fb7f89.js" />
<script src="http://abc.org/2fcab50712467eab4004583eb8fb7f89.js" />
<script src="http://adc.org/2fcab50712467eab4004583eb8fb7f89.js" />
<script src="http://abc.net/2fcab50712467eab4004583eb8fb7f89.js" />

                                    ...
```

Delta - Approach




```
<script src="http://abc.org/2fcab58712467eab4004583eb8fb7f89.js" />
<script src="http://abc.org/2fcab50712467eab4004583eb8fb7f89.js" />
<script src="http://adc.org/2fcab50712467eab4004583eb8fb7f89.js" />
<script src="http://abc.net/2fcab50712467eab4004583eb8fb7f89.js" />

                                    ...
```

⇓

Signature generation

⇓

```
<script src="http://(a(b|d)c.org/2fcab50
                    |abc.org/2fcab58
                    |abc.net/2fcab50)712467eab4004583eb8fb7f89.js"/>
```

Delta - Applications

Detecting infection campaigns:

- New cluster (mostly) malicious?
  - New campaign
- Malicious modification inserted?
  - Campaign spreads
  - Also works when exploit pages are (currently) offline
- Malicious modification removed?
  - End of campaign (potentially)

Delta - Applications

Detecting infection campaigns:

- New cluster (mostly) malicious?
  - New campaign
- Malicious modification inserted?
  - Campaign spreads
  - Also works when exploit pages are (currently) offline
- Malicious modification removed?
  - End of campaign (potentially)

Understanding infection campaigns:

- Same web applications serving malware?
- Same software stack?
- Users with the same browser targeted?
- Users speaking the same language targeted?
- Only users from a set of IP addresses targeted?
- Same shared hosting provider?

Evaluation

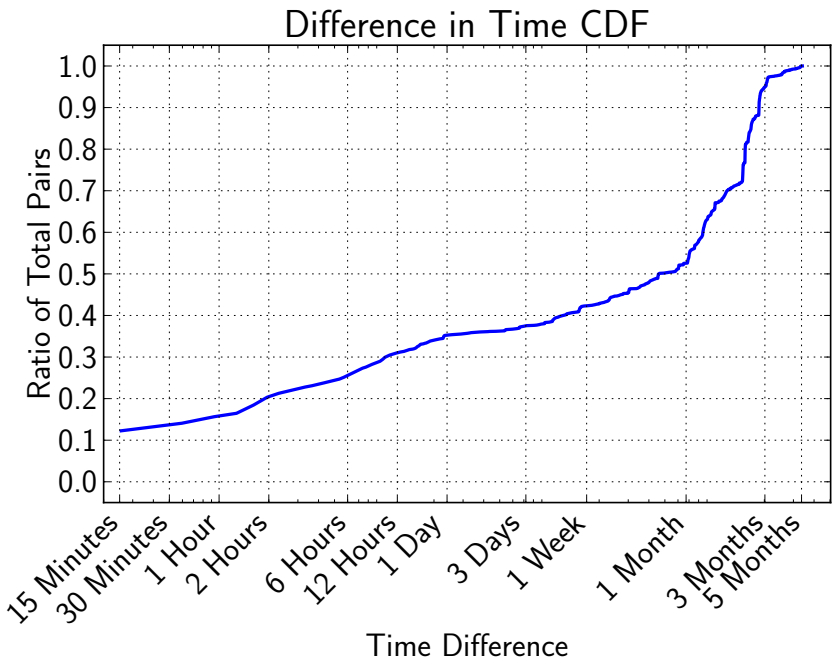
Delta paired with a web crawler

Delta paired with a web crawler

- From January 2013 to May 2013
- Over 12 million unique URLs (max 10 pairs per URL)
- Over 26 million unique pairs of websites (~700GiB)
- Hourly seed: Twitter's trending topics
  - URLs in tweets
  - Yandex's results
- 15 minutes to 1 week recrawl delay

Delta paired with a web crawler

- From January 2013 to May 2013
- Over 12 million unique URLs (max 10 pairs per URL)
- Over 26 million unique pairs of websites (~700GiB)
- Hourly seed: Twitter's trending topics
  - URLs in tweets
  - Yandex's results
- 15 minutes to 1 week recrawl delay



Performance

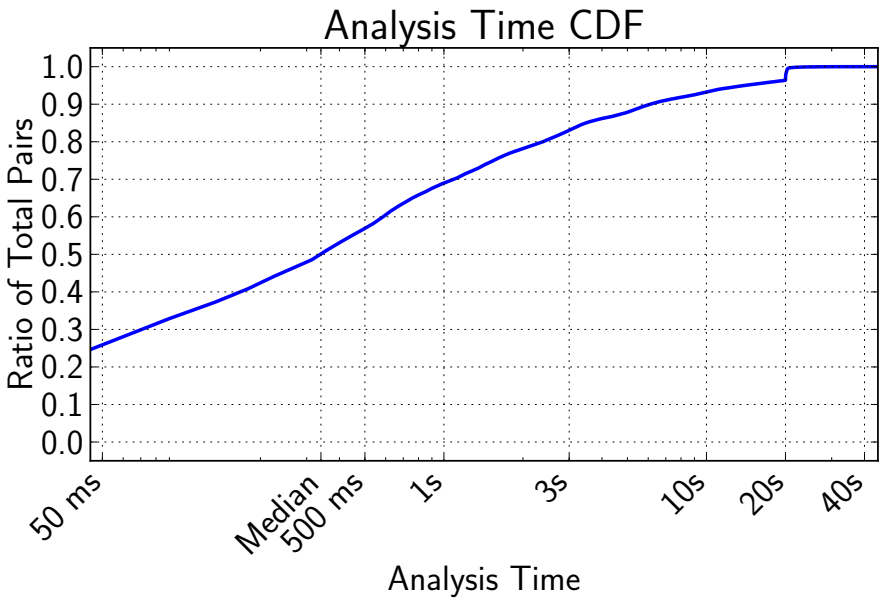
- Viable for large-scale analysis?

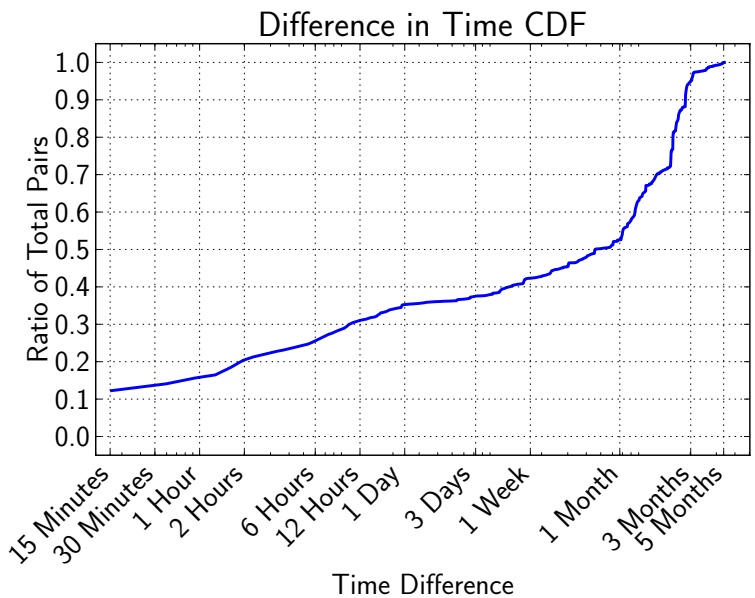
Performance

- Viable for large-scale analysis
- Main bottleneck:
  - HTML Parsing (BeautifulSoup)

Performance

- Viable for large-scale analysis
- Main bottleneck:
  - HTML Parsing (BeautifulSoup)



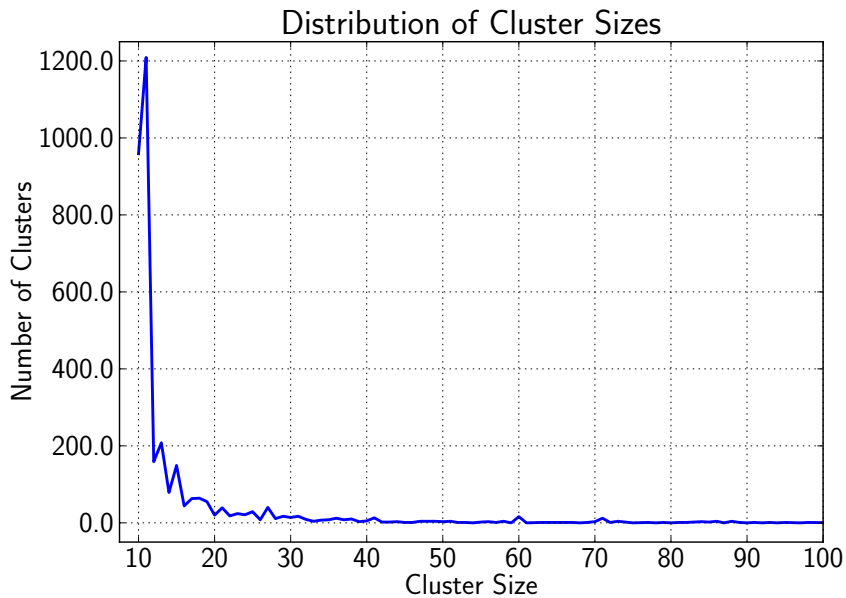


Clusters

- ~67,000 clusters of modifications
- Each cluster has 10 or more observations

Clusters

- ~67,000 clusters of modifications
- Each cluster has 10 or more observations



Cool Exploit Kit

- Redirection to Cool Exploit Kit installation via JavaScript
- Active in April 2013

Cool Exploit Kit

- Redirection to Cool Exploit Kit installation via JavaScript
- Active in April 2013
- 15 websites from 10 unique URLs

Cool Exploit Kit

- Redirection to Cool Exploit Kit installation via JavaScript
- Active in April 2013
- 15 websites from 10 unique URLs
- All Discuz!X (forum software)

Cool Exploit Kit

- Redirection to Cool Exploit Kit installation via JavaScript
- Active in April 2013
- 15 websites from 10 unique URLs
- All Discuz!X (forum software)
- 1 website also included Blackhole

Cool Exploit Kit

- Redirection to Cool Exploit Kit installation via JavaScript
- Active in April 2013
- 15 websites from 10 unique URLs
- All Discuz!X (forum software)
- 1 website also included Blackhole
- Campaign active for over 27 days

Cool Exploit Kit

- Redirection to Cool Exploit Kit installation via JavaScript
- Active in April 2013
- 15 websites from 10 unique URLs
- All Discuz!X (forum software)
- 1 website also included Blackhole
- Campaign active for over 27 days

```
<script type
    ="text/javascript" language="javascript">
  p=parseInt;
  ss=(123) ? String.fromCharCode : 0;
  asgq=" [4036 character obfuscated string] "
       .replace(/@/g,"9").split("!");
  try { document.body&=0.1 } catch(gdsgsdg) {
    zz=3; dbshre=79;
    if(dbshre) { vfvwe=0;
      try { document; }
      catch(agdsg) { vfvwe=1; }
      if(!vfvwe) { e=eval; }
      s="";
      if(zz) for(i=0;i-1374!=0;i++) {
        if(window.document)
          s+=ss(p(asgq[i],16)); }
      if(window.document) e(s); }}</script>
```

Me Gusta

- Found on El Huffington Post
- From January 2013 to May 2013

Me Gusta

- Found on El Huffington Post
- From January 2013 to May 2013
- Nearly 300 website pairs from close to 130 unique URLs

Me Gusta

- Found on El Huffington Post
- From January 2013 to May 2013
- Nearly 300 website pairs from close to 130 unique URLs
- All included Facebook's Like button

Me Gusta

- Found on El Huffington Post
- From January 2013 to May 2013
- Nearly 300 website pairs from close to 130 unique URLs
- All included Facebook's Like button
  with a return link similar to
  `http://www.huffingtonpost.es/2013/04/03/42173.html`

What we have covered:

- Delta approach
  static analysis leveraging web-dynamics to identify unknown infection vectors and support manual analysis
- Practicality
  paired with crawler showed large-scale applicability

Thanks!

Questions?

email kevinbo@cs.ucsb.edu
twitter @caovc
http kevin.borgolte.me